Formation Dataverse

Module Utilisateurs

Martine Barale
CIRAD

Pour citer ce document

Barale M. 2020. Formation Dataverse – Module utilisateurs V7, CIRAD, Montpellier, France, 40 p.

Cadre de réalisation

Ce support de formation a été réalisé dans le cadre de la mise en place et du déploiement au sein du Cirad d'une plateforme de gestion des données de la recherche.

Le projet d'une plateforme de données pour le Cirad a été initié à l'issue des conclusions du groupe de travail *Patrimoine Numérique Scientifique du Cirad* (2014 -2016) qui visait à examiner les sujets liés à la gestion, la valorisation et la diffusion des données de recherche produites par l'établissement dans le cadre de son activité.

Les premières versions de ce support de formation ont été co-produites par Sophie Fortuno et Martine Barale.

cirad

Parcours Numérique Scientifique

Formation

Dataverse
Module utilisateurs

Parcours Numérique Scientifique

Martine Barale (Dist)

Dataverse – Module Utilisateurs

- **Public cible**
  Tout scientifique (chercheur, ingénieur et technicien) amené à **gérer, collecter, produire, partager** et **valoriser** des données.
- **Durée** 1j (7 h)
- **Objectifs de formation**
  Mettre en œuvre de bonnes pratiques en matière de gestion des données :
  - **déposer** et **sécuriser** des jeux de données dans le Dataverse Cirad, les **structurer** et les **documenter** ;
  - appréhender **différents modes d'usages** : du **partage** au sein d'un collectif (projet, UR...) jusqu'à la **diffusion** des données sur Internet ;
  - **rechercher**, **réutiliser** et **citer des données**.

PROGRAMME

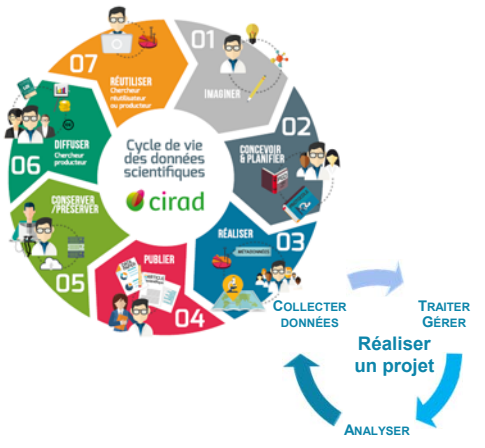
PARTIE I (matin)

- **Séquence 1**
  Le Dataverse du Cirad
- **Séquence 2**
  Stocker, organiser, déposer et partager ses données
- **Séquence 3**
  Préserver, documenter ses jeux de données

PARTIE II (après-midi)

- **Séquence 3**
  (suite et fin)
- **Séquence 4**
  Diffuser, publier ses données
- **Séquence 5**
  Rechercher, réutiliser des données

Programme et périmètre > PARTIE I



Programme

8:30 - 9:00 Accueil, logistique. Tour de table des participants

9:00 - 9:45
Dataverse du Cirad

9:45 - 10:00 - Pause -

10:00 - 11:15
Stocker, organiser, déposer ses données. Les partager

11:15 - 11:45
Préserver et documenter ses données : les métadonnées

11:45 – 13:15
PAUSE REPAS

Programme détaillé **PARTIE I**

- **Séquence 1 – Introduction au Dataverse du Cirad**
  - Cirad et Open data
  - Présentation du Dataverse
  - Principes et enjeux
- **Séquence 2 – Stocker, organiser, déposer, partager**
  - Stocker pour sécuriser
  - Organiser ses données
  - Déposer des fichiers de données
  - Partager ses données
- **Séquence 3 – Préserver et documenter**
  - Décrire ses données par des métadonnées
  - Métadonnées de citation, géospatiales, spécifiques Cirad, disciplinaires
  - Un modèle de saisie : le template

Introduction au Dataverse du Cirad

SEQUENCE 1

Nouveau contexte : l'open data

Des données de recherche ouvertes pour :

- **Capitaliser** sur les résultats précédents
- Encourager la **collaboration**, éviter la dispersion des efforts
- Accélérer l'**innovation**
- Impliquer les **citoyens** et la société
- Et une question de principe reprise dans la loi :

argent public => données publiques

- ► **QUALITE**
- ► **EFFICACITE**
- ► **TRANSFERT**
- ► **TRANSPARENCE**
- ► **RESPECT DU DROIT**

Enjeux pour les chercheurs : entre injonctions…

RESEARCH & INNOVATION Participant Portal H2020 Online Manual
ANR
NERC
wellcome trust
BILL & MELINDA GATES foundation

Incitation ou obligation de partage

REPUBLIQUE NUMERIQUE

Données publiques ?

ELSEVIER
PLOS
WILEY
SPRINGER NATURE
Protocole de Nagoya

Reproductibilité
Preuve
Résultats accessibles à tout public
Préservation
Description et visibilité

… et bénéfices

Augmenter sa visibilité et ses citations
Le partage des données augmente l'attractivité de l'article

Étendre son réseau
Nouvelles opportunités pour échanger/collaborer

Améliorer sa réputation
et attirer sur son profil chercheur

Renforcer ses possibilités de recevoir des financements
(compatibilité avec les exigences des financeurs)

L'ouverture des données : oui mais…

- **C'est compliqué ?…**
  - Utiliser une infrastructure de données
  - Rédiger un plan de gestion des données (DMP)
- **C'est cher ?…**
  - Les frais de stockage, d'organisation et si besoin de dépôt dans un entrepôt (la plupart sont gratuits) sont éligibles dans les appels à projets
- **C'est risqué ?…**
  - La présence de données sensibles (personnelles, secret défense, secret professionnel, secret industriel et commercial, risque pour la protection du potentiel scientifique…) constitue une exception aux exigences d'ouverture des données

« As open as possible, as closed as necessary »

Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020

Cirad et Open data (1)

- *Inventaire Cirad 2017*
  - la majorité des ressources numériques à forte importance stratégique sont **stockées sur des postes de travail** [source PNS – 15 unités pilotes]
- Dimensions EPIC et partenariat au Sud
  - un équilibre à préserver
- En France, la Loi pour une République Numérique est conforme à ce principe pour les données de la recherche
  - **diffusion libre, gratuite et sans restriction** par défaut
  - **sauf exceptions** (droit d'auteur, données personnelles, secret militaire, des affaires, médical, etc.) **ou partenariats**

Cirad et Open data (2)

Au Cirad : Dataverse

Une plateforme pour les données de recherche alliant Open science, sécurisation et partage contrôlé des données

- Lancée en janvier 2018
- Hébergée au Cirad
- **Disponible pour tous les chercheurs Cirad et leurs partenaires dans le cadre de projets communs**

http://dataverse.cirad.fr

Dataverse > En Europe et dans le monde

Entrepôt de données + espace de travail

Plateforme open source supportée par l'université d'Harvard

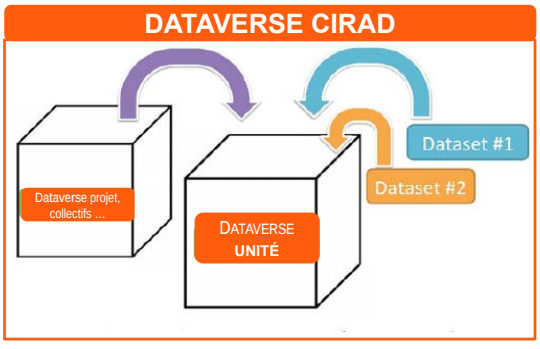
http://dataverse.harvard.edu

Réseau de 52 portails de données scientifiques
(au31/01/2020)

- INRAE Data INRAe https://data.inra.fr/
- IRD DataSuds https://dataverse.ird.fr/
- CDSP Science Po https://catalogues.cdsp.sciences-po.fr/dataverse/archipolis
- CGIAR : presque tous les centres
  http://dataverse.icrisat.org/
  https://data.cifor.org/
  https://dataverse.harvard.edu/dataverse/CIAT
  …

Dataverse > Principe de *containers*

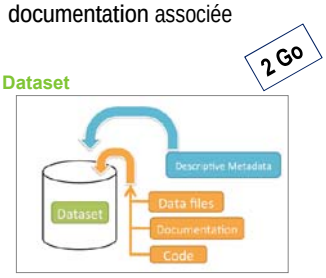
Un **Dataverse** accueille des **Datasets** et d'autres **Dataverses**



Un **Dataset**

✓Est décrit par des métadonnées

✓Reçoit des fichiers de données, du code source et de la documentation associée



Dataverse > *Espaces privé/public (1)*

Entrepôt
Espace privé

- Stocker & décrire (métadonnées)
  - Sauvegarder les fichiers de données (sécuriser)
  - Partager les fichiers (projet, équipe)
- Enrichir les métadonnées
- Documenter
- Soumettre à révision

Portail de diffusion
Espace public Internet

- Diffuser des données
  - obtenir un DOI + citation
  - être visible
- Lier données avec aux :
  - ✓ Publications Agritrop + externes
  - ✓ Datapapers
  - ✓ Thèses
  - ✓ Projets
  - ✓ Sites web, rapports, présentations…

Diffuser

Dataverse > *Espaces privé/public (2)*

Espace de travail

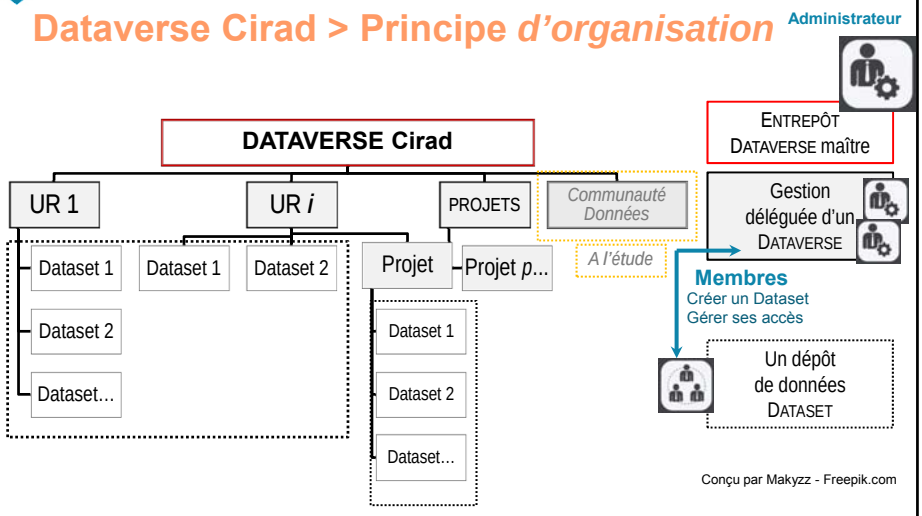
- Gestion des **données de projet / expertise / thèse**
  - Données externes exploitées
  - Nouvelles données produites
- Avant publication :
  - on peut modifier, ajouter, supprimer…

Espace public Internet

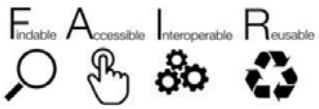
- Diffusion des **données Cirad**
  - Nouvelles données produites par le Cirad et ses partenaires
- \+ **DOI Cirad** (10.18167/…)

= engagement auprès de Datacite de stocker et donner accès durablement à la donnée

Dataverse Cirad > Principe *d'organisation*



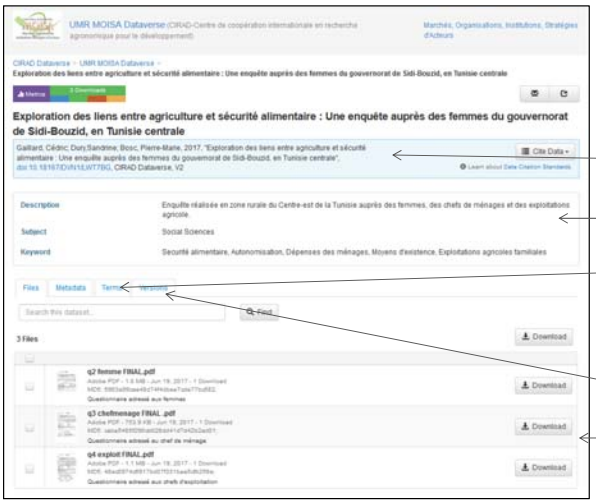
Principes « FAIR »

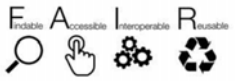


Les données doivent être :

- **F**indable / Faciles à trouver
  Les données sont identifiables et localisables (identifiant unique et pérenne)
- **A**ccessible
  Rendre ces données facilement **A**ccessibles, par les humains et les machines
- **I**nteroperable
  Veiller à ce que les données générées soient **I**nteropérables (formats, vocabulaires, liens) pour permettre le partage et la réutilisation
- **R**eusable / Réutilisables
  Prendre les mesures nécessaires à la **R**éutilisation la plus large possible des données (description riche, licence)

Dataset = « FAIR » Data





Citation de données avec identifiant persistant

Métadonnées

Conditions d'utilisation, licence, accord utilisateur, restrictions

Versions

Fichiers de données (formats ouverts)

Déposer sur le Dataverse : quoi et pourquoi ?

Tous types de données

Les scientifiques décident quelles données déposer :

- **Données brutes**
- **Données traitées**
- **Données analysées**

Pour…

- **Sauvegarder** (règle 3-2-1, activité projet, expertise, formation…)
- **Partager** (projet…)
- **Préserver** (post projet)
- **Diffuser** (livrable projet, embargo, publication...)

IMPORTANT

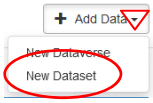
✓ Déposer au plus tôt dans le projet

✓ Veiller à déposer des fichiers de données lisibles et réutilisables dans le temps (pour diffusion ou pas)

Déposer sur le Dataverse : préalable

Pour déposer des données il faut :

1. **Se connecter une première fois**
   - avec son identifiant / mot de passe de messagerie
   - pour créer son compte
2. **Demander au gestionnaire** du Dataverse de votre unité de vous **inscrire dans un groupe** d'utilisateurs
   - Groupe Contributeurs : déposer, gérer ses données, soumettre pour publication
3. **Vous voyez le bouton "Add Data"**
   - Vous pouvez déposer des données



Rôle de la/du/des gestionnaire(s)

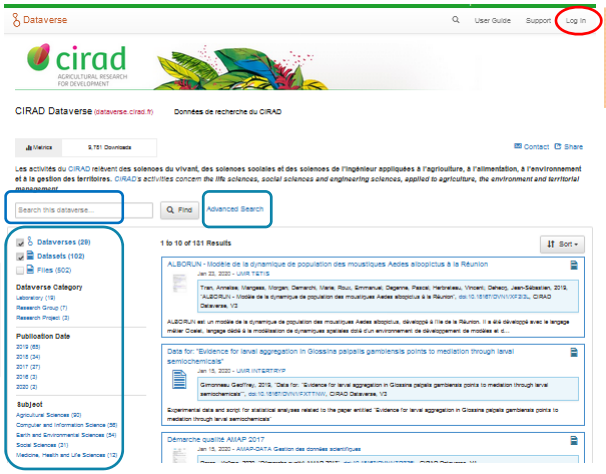
- **Personnalisation du Dataverse** de l'unité : textes de présentation, liens, apparence, contacts...
- **Ajout des nouveaux membres** et définition de leurs droits (voir, déposer, modifier, publier...) en fonction du mode de collaboration choisi par l'unité
- **Création de sous-Dataverses** pour un projet, une thèse, une équipe, etc. ► A réfléchir avant de déposer ⚠
- **Conseil auprès des chercheurs** de l'unité pour leurs premiers dépôts
- **Contrôle des Datasets** qui lui sont soumis (format des fichiers, métadonnées, licence d'utilisation...) et publication (si workflow)
- **Il/elle est formé(e)** : module utilisateur (1 j) et module gestionnaire (atelier de 4 h), il est en relation avec les admin centraux

TP

TP#1 Visite guidée

Données ouvertes diffusées sur Internet
Publication (articles, datapaper...)
Livrable projet…

- Aller sur **http://dataverse.cirad.fr/**



Identification : accès à plus de contenus (selon droits)

Recherche saisie libre

Facettes = Sélection des critères de consultation (sur métadonnées)

Dataset

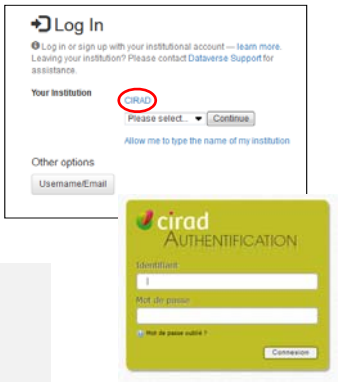
Dataverse

File

TP

TP#2 Se connecter - mon compte - se déconnecter

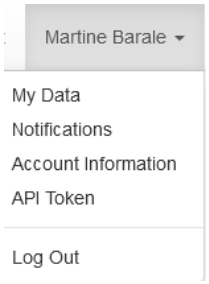
- Connexion sur **http://dataverse.cirad.fr**
- Cliquer sur **Log In** *- en haut à droite -*
- Cliquer sur **CIRAD**
  ou sélectionner CIRAD dans la liste + Continue
- Saisir ses identifiants usuels CIRAD



POUR SE CONNECTER
- Avoir un compte Cirad
- Pour un partenaire ?
  - Voir les modalités de partage de la plateforme
  - Plus d'information sur les comptes informatiques.

TP

TP#2 Se connecter - mon compte - se déconnecter



- Cliquer sur votre identification de connexion

Menu

- **My data** affiche les Dataverses et Datasets auxquels vous avez accès.
- **Notifications** affiche la liste des messages reçus à la suite de la création d'un compte, d'un Dataverse ou d'un Dataset.
- **API Token** génère un code qui vous permet d'utiliser les API Dataverse -non actif-
- **Account Information** affiche les informations de compte, non modifiables quand on est authentifié avec un compte institutionnel CIRAD.
- **Log Out** pour se déconnecter.

Vous avez des données ?

Utilisez le Dataverse du Cirad pour

les stocker, les organiser, les déposer, les partager

SEQUENCE 2

Stocker | Organiser | Déposer | Partager

SOMMAIRE

- Stocker et organiser vos données : quelques rappels
- Déposer des données
- Partager vos données

- ❑ *Avez-vous déjà eu à retrouver des données ou à faire face à une perte de données ?*
- ❑ *Avez-vous eu une incompréhension face à (i) des données produites par un collègue, (ii) des données historiques ?*
- ❑ *Avez-vous besoin de partager de la donnée ?*

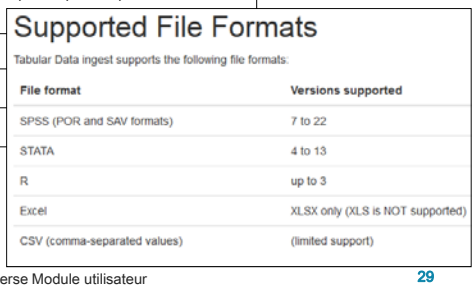
Stocker et organiser vos données : quelques rappels (voir intranet Data)

- **Stocker** pour sécuriser (règle du 3-2-1)
- **Organiser** vos données (dossiers, fichiers)
  - Pour qu'elles soient compréhensibles par les autres
  - Au plus tôt dans le projet (pour ne pas avoir à reformater par la suite)
- **Veiller**
  - Aux règles de nommage (conventions communes)
  - Aux formats ouverts des fichiers
  - A renseigner les propriétés des fichiers (y compris licence)
- **Créer un fichier Readme.txt**
  - Brève description des fichiers et de leurs contenus
  - Organisation des fichiers
  - Dictionnaire des données (nom des colonnes, abréviations, unités…)

Formats ouverts recommandés

Exemples de types de fichiers de données et de formats

Données	Format
Classeur, Feuille de calcul	csv, tsv, xml, ods xlsx, SPSS, STATA
Photo Image	png, tif, tiff
Base de données	csv, txt, xml
Document Script	rtf, pdf/A, odt, xhtml, htm, html, xhtml txt, tex
Video	*ogg*
Audio	mp3
Compressé	zip, tar



Supported File Formats

Tabular Data ingest supports the following file formats:

File format	Versions supported
SPSS (POR and SAV formats)	7 to 22
STATA	4 to 13
R	up to 3
Excel	XLSX only (XLS is NOT supported)
CSV (comma-separated values)	(limited support)

Déposer des fichiers de données (1)

Créer un Dataset = ensemble cohérent de fichiers décrits par les même métadonnées.

- **5 métadonnées** nécessaires (obligatoires)

Métadonnée	Description
Title	Titre en langue originale
Author	Nom, Prénom (ou Organisation pérenne)
	Affiliation : CIRAD, UR (acronyme ou intitulé court), pays d'affectation
	ORCID : n° ORCID de l'auteur si il existe
Contact	Nom, Prénom du destinataire des demandes d'informations ou d'accès
	Affiliation : CIRAD, UR (acronyme ou intitulé), pays affectation
	Email : de type contact@cirad.fr
Description	Résumé bref décrivant l'objet, la nature et la portée des données, le lieu et la couverture temporelle.
Subject	Au-moins une grande discipline. Exemple : Agricultural Sciences

- un doi est pré-affecté dès la validation de la création du Dataset

Déposer des fichiers de données (2)

Téléverser les fichiers

- **Formats**

  Tous les formats sont acceptés.

  En fin de projet veiller à déposer des formats ouverts pour que vos fichiers soient lisibles et réutilisables dans le temps.

- **Facilités de visualisation des contenus**

  RData, SPSS, STATA, CSV, XLSX seulement ***(xls non supporté)***

- **Taille du Dataset limitée à [2 Go]** (1 à 100 fichiers)

- **Fichier compressé**

  Un **fichier zip** est automatiquement décompressé dans le Dataset.

  Les **autres formats de fichiers compressés (tar.gz**,…) restent dans leur format d'origine sur la plateforme.

TP

TP#3 Créer un Dataset

1. Connexion sur **http://dataverse.cirad.fr**
2. Aller sur le Dataverse dans lequel vous voulez créer un Dataset.
3. Cliquer sur **Add Data** > **New Dataset**
4. Saisir les **5 informations obligatoires** (*)
5. **Téléverser** les fichiers **> Select Files to Add**

6. Valider la **création** du Dataset

Verrouiller les fichiers déposés (via sélection)

- **Edit Files > Restrict** Saisir les **Terms of Access** pour en préciser la raison *: projet en cours, embargo, données réservées à usage interne …*
- **Edit Files > Unrestrict** Déverrouille

Partager

PARTAGER SES DONNÉES

A. Tiers sans compte Cirad
   - Cas du reviewing d'un article soumis
   - Cas du partage complet avec un tiers

B. Tiers avec un compte Cirad
   - Permissions sur le Dataset

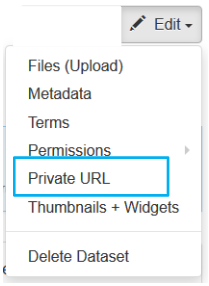
Partager ses données > Tiers SANS compte Cirad

A. Partager > URL privée

- pour « **reviewer** » **ponctuellement** un Dataset
  - soumission publication –
- pour partager **ponctuellement** : les **métadonnées en lecture** ET totalement les fichiers d'un Dataset avec un tiers - projet - (au-delà des restrictions posées)

Dataverse permet de **créer une URL privée** (lien)
à envoyer par mail à un tiers ou à la revue scientifique

Au niveau du Dataset



Important : penser à désactiver l'URL privée

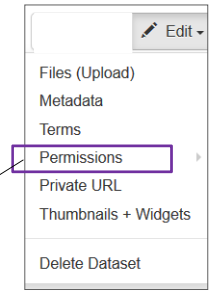
Partager ses données > Tiers AVEC un compte Cirad

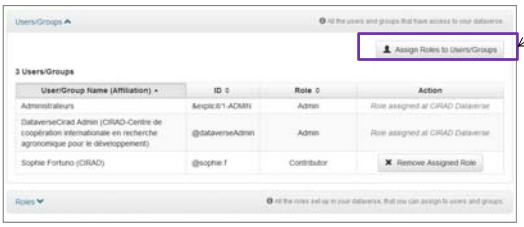
B. Dataset > Permissions

- pour partager **durablement** un Dataset avec des tiers
  - projet, collaboration, équipe –
- Le Dataset est visible et accessible

Bouton **Assign Roles to Users/Groups** : attribution d'un *rôle* particulier à un utilisateur *(ou groupe d'utilisateurs)*

Au niveau du Dataset





Partager ses données > Tiers AVEC un compte Cirad

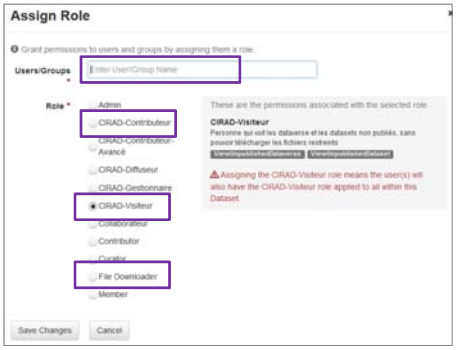
C. Dataset > Permissions > Assign Roles

1/ Saisir le nom de l'utilisateur

a) Accès en lecture aux métadonnées du Dataset + Téléchargement des fichiers, **si non restreints**.
Choisir le rôle CIRAD-Visiteur

b) Accès en lecture aux métadonnées du Dataset + Téléchargement des fichiers, **même si restreints**
Choisir 2 rôles (itérativement)
CIRAD-Visiteur
File Downloader

c) Accès en **modification sur le Dataset**
Choisir 2 rôles (itérativement)
CIRAD-Visiteur
CIRAD-Contributeur



TP

TP#4 Partager un Dataset

- Aller sur le Dataset créé

Exercice 1
- Créer une URL privée
- L'envoyer à votre binôme par mail

Exercice 2
1. Créer une permission pour que votre binôme ait accès au Dataset en consultation (métadonnées) avec téléchargement des fichiers non restreints

2. Créer une permission pour que votre binôme ait accès en modification au Dataset

Préserver et documenter ses données

SEQUENCE 3

Décrire ses données : pourquoi ?

- ❑ *Avez-vous déjà recherché des données dans un entrepôt ?*
- ❑ *Quels étaient vos critères ?*
- ❑ *Quelles difficultés pour trouver des données pertinentes ?*

- Pourquoi documenter ses données ?
  - Pour qu'elles soient **retrouvées, comprises et réutilisées**
  - Critères de recherche = **métadonnées**
    - ⇨ Titre
    - ⇨ Pays de collecte
    - ⇨ Dates
    - ⇨ Sujets
    - ⇨ Types de données, …
  - **Accessibilité, droits de réutilisation** : pour qui ? pour faire quoi ?

Les métadonnées

- Description d'un jeu de données dans le Dataverse Cirad
  1) Métadonnées de **citation** : titre, auteur, description, …
  2) Métadonnées **géospatiales**
  3) Métadonnées **spécifiques Cirad** : filières, thématiques
  4) Métadonnées **disciplinaires** : SHS, Life sciences…
- Utilisation de « standards » de métadonnées et de vocabulaires
- Recommandations de saisie = le « **template** »
  - Formats de saisie homogènes (auteurs, affiliations…)
  - Utilisation de référentiels (topic, mots-clé)
- Intérêt et enjeux
  - **Interopérabilité** des entrepôts
  - **Qualité et homogénéité** des métadonnées (recherche, facettes…)

1) Les métadonnées de citation

- **Norme Dublin Core (DC)** = ISO 15836
  largement utilisée au niveau international

  - 15 champs ± champs spécifiques (DC étendu)

- Titre (dc:title)
- Sujet (dc:subject)
- Description (dc:description)
- Origine de l'information (dc:source)
- Langue (dc:language)
- Relation avec d'autres ressources (dc:relation)
- Couverture chronologique et géographique (dc:coverage)
- Auteur (dc:creator)
- Contributeur (dc:contributor)
- Editeur (dc:publisher)
- Conditions d'utilisation (dc:rights)
- Date (dc:date)
- Type de document (dc:type)
- Format (dc:format)
- Identifiant (dc:identifier)

1) Les métadonnées de citation (suite)

- **5 champs obligatoires** dans le Dataverse Cirad
  - Titre (complet, informatif)
  - Auteur(s) = nom (± affiliation, identifiant)
  - Contact = email (± nom, affiliation)
  - Description (texte libre)
  - Sujet (liste de 13 disciplines scientifiques – Harvard)

    ➢ **On obtient une citation « propre » et un DOI**
    (identifiant numérique pérenne du jeu de données)

- **Enrichir la description au maximum** (on peut y revenir)
  - ⇨ **Types de données**, dates, contributeurs, langue, publications liées, etc.
  - ⇨ Pour faciliter le repérage, la compréhension et la **réutilisation**

2) Les métadonnées géospatiales

- 3 métadonnées
  - ✓ coverage : pays +/- province ou état, ville…
  - ✓ unit = niveau de la collecte des données (un district, un village…)
  - ✓ bounding box : latitude / longitude – coordonnées GPS (WGS 84)

X (nord, ouest)

(sud, est) X

Geospatial Metadata	
Geographic Coverage	Cameroon Littoral Njombé
Geographic Bounding Box	9.38'49 4.34'12

Geospatial Metadata	
Geographic Coverage	Madagascar Itasy Madagascar Sofia
Geographic Unit	Districts

3) Les métadonnées spécifiques Cirad

- Selon classifications SIRH (idem CV…)
  - 21 **filières**
  - 19 **thématiques scientifiques**

Cirad Metadata	
Filière(s)	Travail; Volailles; Camélidés; Petits ruminants; Bovins, zébus, buffles, yack; Plantes diverses; Légumineuses fourragères; Légumineuses alimentaires; Agrumes; Autres céréales
Thématique(s)	Sécurité alimentaire; Déserts et zones arides; Elevage et produits animaux (filières animales)
Autre(s) thématique(s)	Empowerment

Cirad Metadata	
Filière(s)	Bananier et plantain
Thématique(s)	Sécurité alimentaire; Ecosystèmes cultivés tropicaux; Plantes et produits tropicaux (filières végétales)

Vers une interopérabilité entre les outils Cirad…

4) Les métadonnées disciplinaires

- **Différents standards** proposées (ou pas) dans les Dataverses d'unité (selon disciplines pertinentes)
  - ⇨ Paramétrage par le gestionnaire du DV d'unité
- Non obligatoires… mais **très importantes** pour
  - ⇨ Permettre une bonne compréhension du jeu de données
  - ⇨ **Favoriser la réutilisation** par les pairs
- Actuellement **deux standards** intéressants pour le Cirad
  - ⇨ **Social science and humanities** = données d'enquête (standard DDI)

  21 métadonnées proposées
  - ⇨ **Life sciences** = types d'essai, organismes… (standard ISA-Tab)

  9 métadonnées proposées
  - ⇨ A prévoir : l'intégration d'autres standards disciplinaires par Dataverse.

cirad

Parcours Numérique Scientifique

Formation

Module
Dataverse - utilisateurs

(2° partie)

Programme et périmètre > PARTIE II

Programme

13:15 - 14:00
Préserver et documenter ses données : Template de saisie

14:00 - 15:00
Diffuser, publier ses données

15:00 - 15:15 - Pause -

15:15 - 16:15
Rechercher, réutiliser des données

16:15 - 16:30
Evaluation de la formation

Programme détaillé PARTIE II

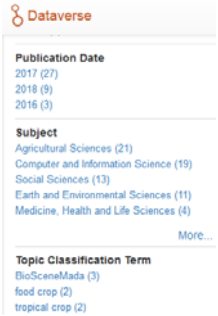
- **Séquence 3 – Préserver et documenter**
  - Décrire ses données par des métadonnées
  - Métadonnées de citation, géospatiales, spécifiques Cirad, disciplinaires
  - Un modèle de saisie : le template
- **Séquence 4 – Diffuser, publier**
  - Niveaux d'accès aux métadonnées, aux fichiers
  - Conditions d'utilisation et licences
  - Citer ses données, publier un data paper
- **Séquence 5 – Rechercher, réutiliser**
  - Rechercher dans le Dataverse Cirad
  - Rechercher dans d'autres entrepôts
  - Citer correctement, respecter les droits d'utilisation

Le modèle de saisie ou « template »

- Des recommandations pour la saisie
  - **Intégrées dans le formulaire de saisie** des md de citation + géospatiales
    - ⇨ À effacer et remplacer par vos informations
- Pour inciter à l'adoption de « **bonnes pratiques** » au Cirad
  - ⇨ Le « producer » est une institution (le plus souvent le Cirad)
  - ⇨ Le mail de contact doit normalement être un mail@cirad.fr, …
- Pour **homogénéiser le format** de certaines données
  - ⇨ Les dates, les noms de personnes (auteurs, contact…), les affiliations, …
- Pour favoriser l'utilisation de **référentiels et de vocabulaires contrôlés**
  - ⇨ Identifiant chercheur : ORCID
  - ⇨ Lien vers un document : DOI
  - ⇨ Mots-clés, classification thématique : référentiels FAO (Agrovoc, Agris).

Avantages (et inconvénient) du « template »

- favoriser l'**interopérabilité** entre les entrepôts
- garantir l'homogénéité et la **cohérence des données**
- faciliter la **recherche via les facettes**
  - ⇨ Topic classification (thématiques Agris / FAO) ⟹
  - ⇨ Keywords (descripteurs Agrovoc / FAO)
  - ⇨ Producer name (Cirad ou autre institution), …

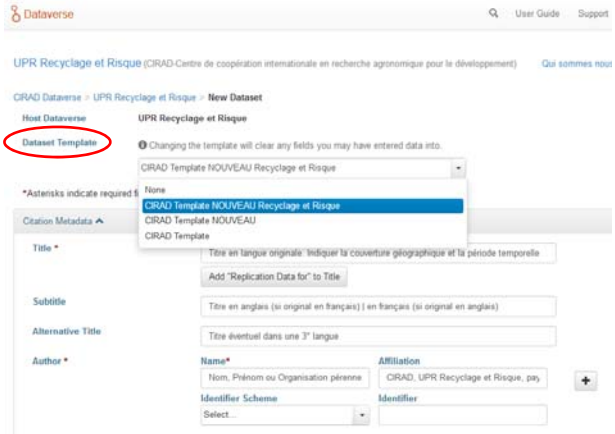


- **aider les déposants** « débutants » en proposant des **listes indicatives ou des exemples de contenus** pour certains champs en saisie libre
  - ⇨ Exemple : Kind of data

⚠ **Penser à supprimer les informations du template avant de publier le Dataset**

Utiliser le « template »

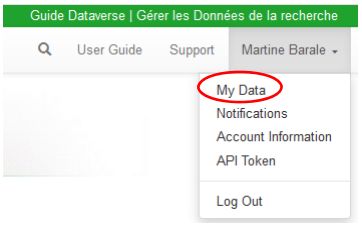
- Plusieurs templates peuvent être proposés dans un Dataverse d'unité
- On a le choix d'utiliser ou non un template



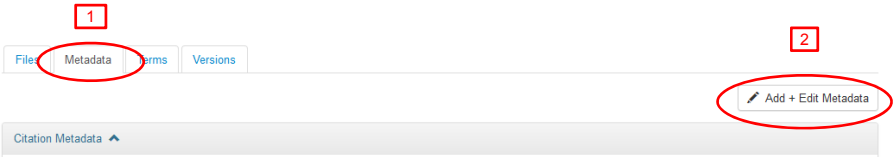
TP

TP#5 Complétez la description de vos données

- Revenir sur votre jeu de données



- Onglet Metadata
- Cliquer sur « Add + Edit Metadata »



Diffuser des données, les publier

SEQUENCE 4

Pourquoi diffuser ses données ?

- ❑ *Participez-vous à un projet pour lequel un des bailleurs impose ou recommande la diffusion des données ?*
- ❑ *Etes-vous confronté à différentes situations pour la diffusion de vos données : confidentielles, partageables au sein d'un groupe, diffusables sur Internet ?*
- ❑ *Vous est-il arrivé de renoncer à diffuser vos données par crainte de l'utilisation qui pourrait en être faite ?*

- Rendre publics vos jeux de données de recherche, c'est
  - assurer leur **visibilité**,
  - les partager et être crédité de leur **paternité**,
  - faciliter leur réutilisation et donc une meilleure **valorisation**.
- Dans certains cas, la diffusion est obligatoire !

Diffuser ses données : quand, comment ?

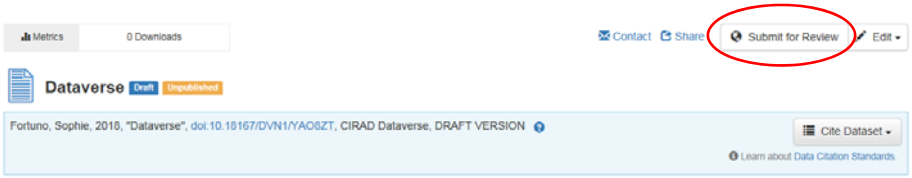
- **Quand faut-il diffuser ses données ?**
  - **A la fin du projet**, selon les préconisations du PGD ou du contrat

  [A défaut : revenir vers les partenaires du projet pour accord]
  - Pendant / **après la parution** des articles, data papers, livrables du projet
  - On peut prévoir un **embargo** sur l'accès aux fichiers
  - Autre choix possible : **jamais** (conservation d'un patrimoine).
- **Que faut-il faire pour diffuser ses données ?**
  - Définir le **niveau d'accès** souhaité aux métadonnées, aux fichiers
  - Définir les conditions d'utilisation et choisir une **licence** **(onglet Terms)**
  - Rendre le jeu de données public : bouton **Publish** (si droits)
  - **Ou : soumettre** le jeu de données pour « publication »
  - Pour reviewing (article, data paper) : fournir une « **private url** ».

Diffuser ses données

- **Si on détient les droits de « publication »** dans le DV de son unité

- **Si la « publication » est modérée** dans le DV de son unité



Diffuser ses données : points de vigilance

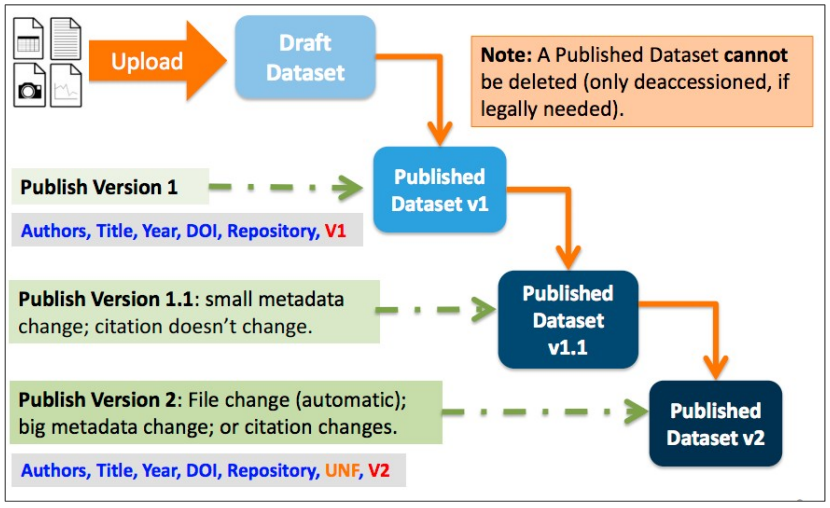
- La publication du jeu de données est définitive
  - On peut faire des modifications ► nouvelle version
  - L'attribution du DOI engage le Cirad auprès de Datacite

Deaccession Your Dataset [not recommended]

Deaccessioning a dataset or a version of a dataset is a very serious action that should only occur if there is a legal or valid reason for the dataset to no longer be accessible to the public.

- Avant de publier, vérifiez les métadonnées et supprimez tous les éléments du template
- Vérifier la présence de la licence appropriée

Dataset > Principe de montée de *version*



Diffuser ses données en libre accès ? (1)

- Définir le **niveau d'accès** au jeu de données en fonction
  - Des obligations légales si **données publiques**
  - Des incitations ou **obligations des bailleurs**. Ex : H2020
    « As open as possible, as closed as necessary »
  - Des politiques des **éditeurs de revues** (si article prévu)
  - Des **partenariats** : contrat de consortium, de coproduction
  - Des contraintes spécifiques du projet et des règles définies dans le **PGD**

Diffuser ses données en libre accès ? (2)

- **RAPPEL -** Doivent faire l'objet d'une attention particulière :
  - les **données à caractère personnel** (nom, prénom, adresse de messagerie, adresse postale, … dans le cadre ***d'enquêtes, colloque, liste de contacts)***
  - les **données relevant de la propriété intellectuelle et industrielle**
  - les **données sensibles**
- **Anonymisation ou pseudonymisation** : quelques ressources
  - Outil **Amnesia** (projet H2020) : https://amnesia.openaire.eu/index.html
  - Guide SGMAP/AGD : https://github.com/SGMAP-AGD/anonymisation/wiki
  - Données personnelles : anonymisation ou pseudonymisation ? par Charlotte Galichet, Avocat : https://www.village-justice.com/articles/donnees-personnelles-anonymisation-pseudonymisation,26194.html

Diffuser ses données en libre accès ? (3)

- Différencier le niveau d'accès aux métadonnées / aux fichiers
  - **Métadonnées ouvertes** : pour faire connaitre le jeu de données
  - **Fichiers : trois modalités d'accès possible** (téléchargement)
    - ⇨ Accès libre (par défaut)
      - – on peut demander à l'utilisateur de préciser l'usage prévu
    - ⇨ Accès sur demande [Request Access]
      - – Peut permettre de sélectionner les utilisateurs / utilisations
      - – Peut générer des partenariats (réutilisations conjointes)
    - ⇨ Pas d'accès
- On peut faire évoluer le niveau d'accès tout au long du projet
  - **Limité** à l'équipe / l'unité pendant le projet (**données primaires, provisoires**)
  - Accès ouvert aux métadonnées en attendant les publications (**embargo**)
  - **Accès ouvert** aux fichiers, ou **sur demande**, après valorisation.

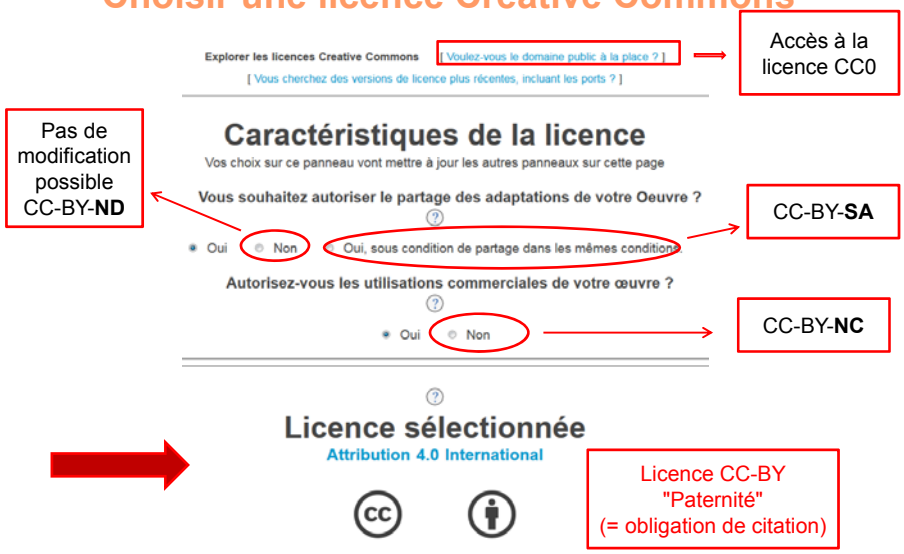
Diffuser ses données : pour quelle utilisation ?

- Définir les **conditions d'utilisation** et les restrictions éventuelles
  - Utilisation commerciale / modification / intégration à d'autres données… ?
  - Citation de l'auteur ? (exigence "classique" mais limite pour la fouille de texte)
  - Rappel : pas de restriction pour les données publiques 
- Utiliser les **licences ouvertes**
  - Pour **faire connaitre les usages autorisés** aux utilisateurs potentiels
  - **Contrats** standards, sans contact entre les parties, qui s'imposent aux chercheurs souhaitant réutiliser les données
  - Plusieurs licences disponibles
- Autres possibilités
  - Texte court précisant les obligations et interdictions
  - Si licence spécifique, charte d'usage… : lien vers site du projet ou contact.

Les licences ouvertes

- Les licences Creative Commons (CC)
  - **6 licences** correspondant à des combinaisons de droits / restrictions
  - Formulaire de **sélection en ligne** : j'accepte / je refuse pour chaque droit
  - On choisit une licence et on copie/colle le code obtenu dans *Terms of use*
  - Par défaut, Dataverse propose **CC0 – Public domain dedication**
    - ⇨ Très large : toutes utilisations autorisées, aucune restriction
    - ⇨ Pas de citation des auteurs
  - On peut refuser CC0 (template) et indiquer une autre licence CC au choix
  - Par défaut : «Utilisation soumise à l'autorisation de l'auteur ou du Cirad »
- Autres licences
  - Licence ouverte Etalab (données publiques françaises – équivalent CC-BY)
  - ODbL (bases de données)…

Choisir une licence Creative Commons



Accès à la licence CC0

Pas de modification possible CC-BY-**ND**

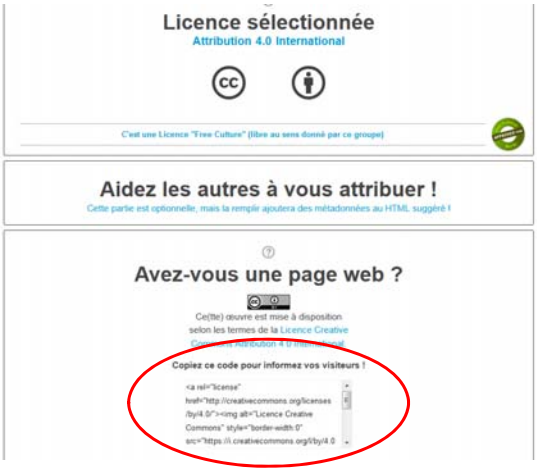
CC-BY-**SA**

CC-BY-**NC**

Licence CC-BY "Paternité" (= obligation de citation)

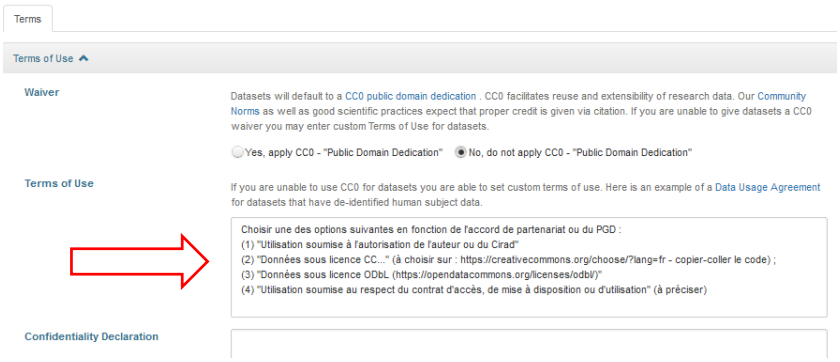
Insérer la licence Creative Commons (1)

- Copier le texte associé à la licence choisie



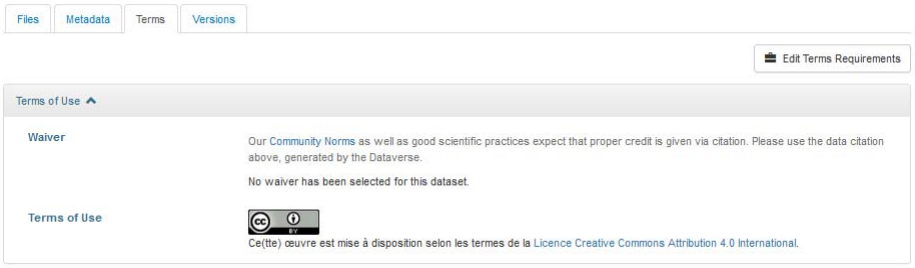
Insérer la licence Creative Commons (2)

- Coller le texte dans les "Terms of use" de votre Dataset

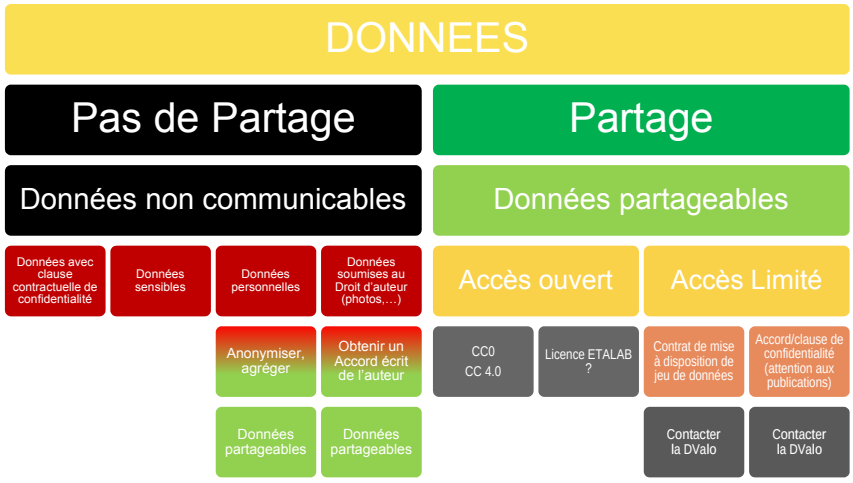


Insérer la licence Creative Commons (3)

- Le logo de la licence est inséré dans le Dataset



Diffuser ses données : en résumé



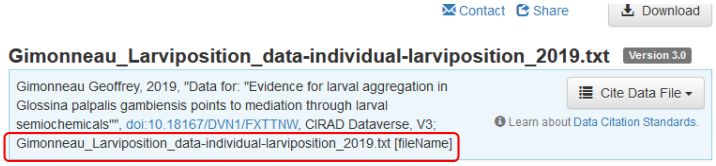
Citer vos données dans vos articles

- Utilisez la **citation et** le **DOI** fournis par le Dataverse

- Transmettez le **lien DOI** plutôt que l'URL longue

http://dx.doi.org/10.18167/DVN1/FXTTNW

- Si nécessaire on peut **citer un fichier particulier** du Dataset



Citer et publier ses données

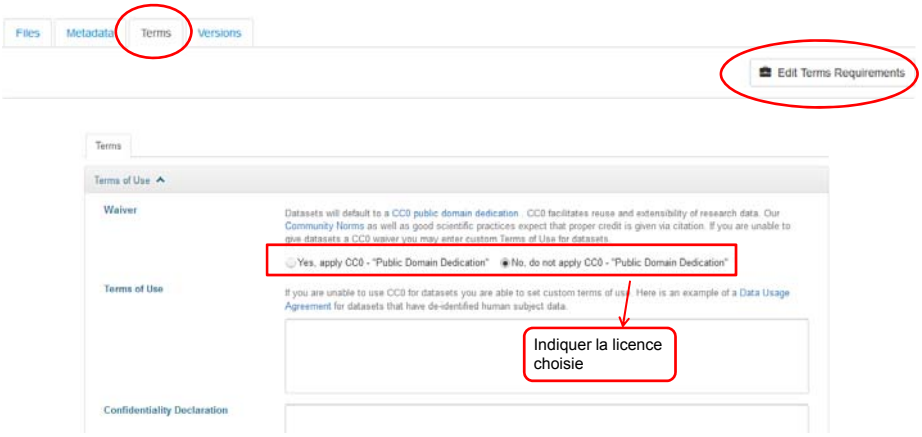
- **Citez vos données dans vos articles scientifiques**
  - Utilisez la **citation,** le **DOI** et le lien DOI fournis par le Dataverse
  - Si nécessaire on peut citer un fichier particulier du Dataset
  - De plus en plus de revues demandent que les données soient accessibles

- Un + : **publiez un data paper** pour faire connaitre vos données
  - Type d'article particulier décrivant un jeu de données, publié dans une revue classique ou spécifique (data journals)
  - Le jeu de données doit être **déposé dans un entrepôt** et accessible
    - ⇨ L'accès aux fichiers peut être différé (embargo)
    - ⇨ Description complète dans le Dataverse facilitera la rédaction du data paper
  - Avantages :
    - ⇨ Met en avant le **potentiel de réutilisation** des données
    - ⇨ Génère des **citations** et des partenariats
    - ⇨ Renforce votre **crédibilité** en tant que chercheur (transparence, intégrité sc)

TP

TP#6 Complétez votre dépôt : onglet Terms

- Onglets « Terms » puis « Edit Terms Requirements »



Rechercher et réutiliser des données

SEQUENCE 5

Réutiliser des données

- ❑ *Si vous avez déjà réutilisé des données, où et comment les avez-vous obtenues ?*
- ❑ *Aviez-vous une autorisation explicite des auteurs, ou avez-vous pensé « si elles sont sur Internet, c'est qu'on peut en faire ce qu'on veut » ?*
- ❑ *Aviez-vous connaissance de restrictions d'usage, par exemple ne pas les modifier, ne pas les utiliser pour une étude rémunérée… ?*

- Réutiliser des données de recherche, c'est
  - Les citer correctement,
  - Prendre connaissance des conditions d'utilisation (licence),
  - Identifier les restrictions et les respecter.

Rechercher des données dans le Dataverse

- **Recherche simple** = cartouche type Google
  - Saisir des mots significatifs : mots du titre, nom d'auteur, mots-clé…
  - Exemple : banane guadeloupe
- **Recherche avancée** = par champ de métadonnées
  - Exemple : « Description text » = banane / « Publication date » = 2017
- **Sélection à partir des facettes** (restrictions itératives)
  - Choix du type de résultat recherché : Dataverses, Datasets, fichiers
  - Consultation des contenus du Dataverse à partir d'un sujet, un thème, un mot-clé, un pays…
  - Restriction d'un résultat par type de données, langue, année, producteur…

Rechercher dans d'autres entrepôts

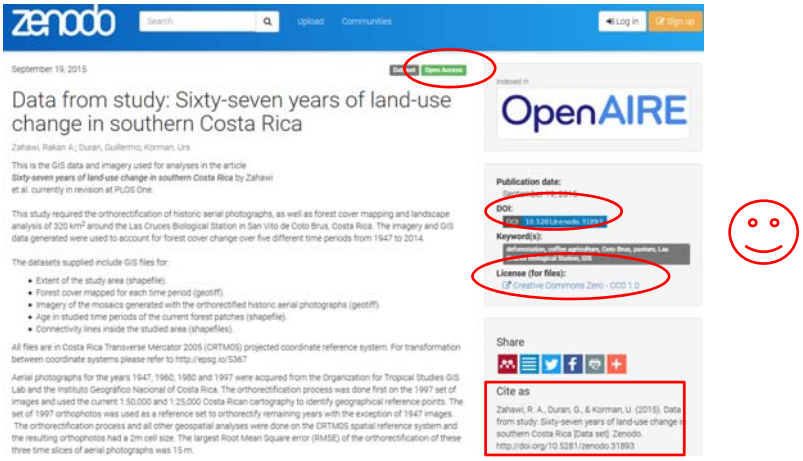
- **Trouver un entrepôt de données** : le site **re3data.org**
  - REgistry of REsearch Data REpositories
  - Découverte par sujet, type de contenu, pays
  - Recherche par mots ou sélection sur listes

- **Rechercher dans Zenodo** (UE - Cern, accès libre et gratuit)

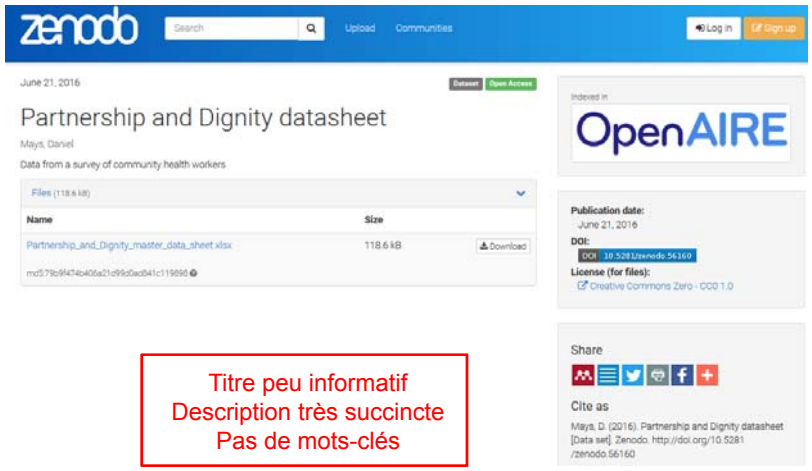
  https://zenodo.org/
  - Exemple : agriculture – 11859 résultats dont 471 jeux de données
  - Exemple : "climate change" – 2240 résultats dont 162 jeux de données

- Autre source : **Data Citation Index** (Clarivate Analytics)
  - payant, accessible au Cirad (intranet Dist > Revues ebooks bases de données)

Un exemple de jeu de données dans Zenodo



Un autre exemple…



Titre peu informatif
Description très succincte
Pas de mots-clés

77

TP

TP#7 Rechercher dans Zenodo

- Recherchez dans Zenodo des jeux de données dans votre thématique
- Les analyser, évaluer leur description
- Identifier les conditions d'utilisation / licences

- On peut aussi faire une recherche dans Data citation index

Aide et supports

- **Cirad**
  - Site intranet DATA intranet-data.cirad.fr
  - Guides utilisateurs https://intranet-data.cirad.fr/outils-et-ressources
  - Site coopIST Gérer des données
- **Dataverse**
  - Dataverse http://dataverse.cirad.fr dataverse@cirad.fr
  - Guide utilisateur Guide utilisateur Dataverse
  - Site intranet DATA intranet-data.cirad.fr rubrique Dataverse
- DORANum http://doranum.fr/
  - Ressources thématiques, tutoriels, fiches synthétiques, quizz, vidéos…
- **Creative Commons** https://creativecommons.org/choose/?lang=fr
  - Formulaire pour le choix d'une licence Creative Commons

MERCI DE VOTRE ATTENTION

Pour toute information :

dataverse@cirad.fr